D368–D371 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 4 November 2007*
doi:10.1093/nar/gkm833

Glycoconjugate Data Bank:Structures—an annotated glycan structure database and *N*-glycan primary structure verification service

Taku Nakahara[1], Ryo Hashimoto[1,4], Hiroaki Nakagawa[2], Kenji Monde[1], Nobuaki Miura[3] and Shin-Ichiro Nishimura[1,*]

[1]Laboratory of Advanced Chemical Biology, [2]Laboratory of Glyco-Fine Chemistry, [3]Sun Microsystems Laboratory of Computational Molecular Life Science, Graduate School of Advanced Life Science, Frontier Research Center for the Post-Genomic Science and Technology, Hokkaido University, Sapporo 001-0021 and [4]Science & Technology Systems Inc. Tokyo, 150-0002, Japan

Received August 13, 2007; Revised September 20, 2007; Accepted September 21, 2007

ABSTRACT

Glycobiology has been brought to public attention as a frontier in the post-genomic era. Structural information about glycans has been accumulating in the Protein Data Bank (PDB) for years. It has been recognized, however, that there are many questionable glycan models in the PDB. A tool for verifying the primary structures of glycan 3D structures is evidently required, yet there have been no such publicly available tools. The Glycoconjugate Data Bank:Structures (GDB:Structures, http://www.glycostructures.jp) is an annotated glycan structure database, which also provides an *N*-glycan primary structure (or glycoform) verification service. All the glycan 3D structures are detected and annotated by an in-house program named 'getCARBO'. When an *N*-glycan is detected in a query coordinate by getCARBO, the primary structure of the glycan is compared with the most similar entry in the glycan primary structure database (KEGG GLYCAN), and unmatched substructure(s) are indicated if observed. The results of getCARBO are stored and presented in GDB:Structures.

INTRODUCTION

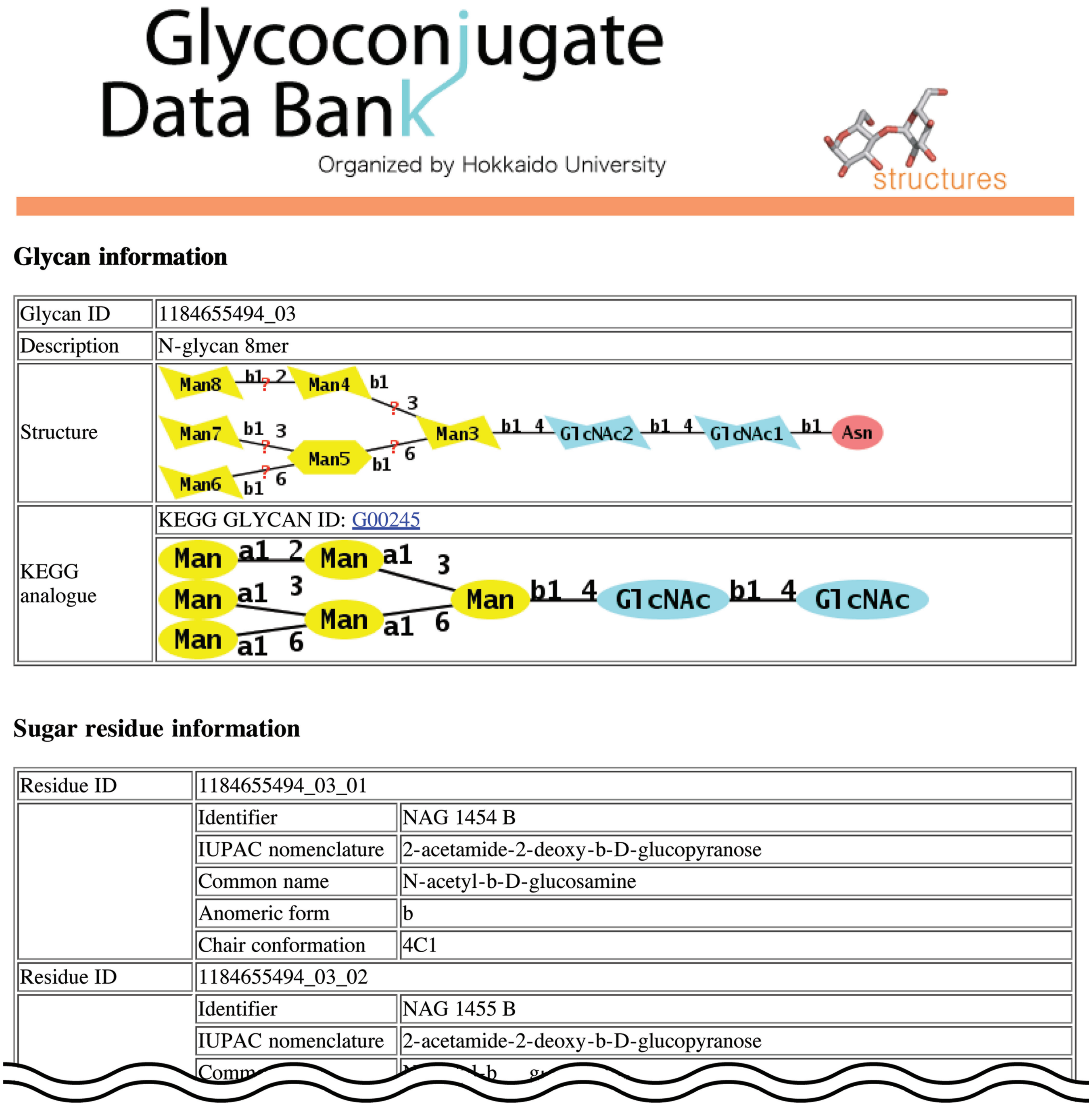
Glycans (carbohydrate chains) are one of the major classes of biological molecules, like nucleic acids and proteins. There is a substantial amount of structural information about glycan 3D structures in the Protein Data Bank (PDB) (1), and this is the major resource in the structural biology of glycans. Most of these glycan structures coexist with proteins as ligands or modifications of glycoproteins. There are two major classes of modified glycans, namely *N*-glycans (asparagine-linked glycans) and *O*-glycans (serine or threonine-linked glycans).

The accuracy of the glycan structures in the PDB has been in question for years. It was reported that there are some hundreds of asparagine-linked *N*-acetylglucosamines with inaccurate anomeric forms (2,3). Furthermore, ~30% of saccharide units in the PDB contain errors, mainly in monosaccharide nomenclature (4). So far, there is only one web service for checking the correspondence between the structure and nomenclature of a saccharide unit in a PDB file (5).

Recently, Crispin *et al.* (6) pointed out that glycan structures with unreported primary structure motifs were observed in the PDB. Models of glycan 3D structures built by X-ray crystallography should be supported by the glycan primary structure data derived by other methods (NMR, HPLC, mass spectrometry and so on). Our recent comprehensive survey revealed that 13.6% of glycans in the PDB contains substructure(s) which are not found in the glycan primary structure database (7), which suggested that the PDB contains a substantial number of glycans of which primary structures cannot be synthesized by known biological pathways. A tool to verify the glycan primary structures of 3D models is in urgent demand.

The Glycoconjugate Data Bank:Structures (GDB: Structures) is a database of annotated glycan structures. Glycans in the PDB were detected and annotated by a computer program named 'getCARBO'. Furthermore, *N*-glycans were compared with entries in the glycan primary structure database, KEGG GLYCAN (8). The KEGG GLYCAN entry with the most similar structure to the PDB glycan is selected by using a glycan structure search function implemented in getCARBO, and unmatched substructure(s) between the glycan found in the PDB and the most similar KEGG GLYCAN entry were noted graphically. GDB:Structures also provides a web service for checking glycan structures in a user-uploaded

*To whom correspondence should be addressed. Tel: +81 11 706 9043; Fax: +81 11 706 9042; Email: shin@glyco.sci.hokudai.ac.jp



Glycoconjugate Data Bank

Organized by Hokkaido University

structures

Glycan information

Glycan ID	1184655494_03
Description	N-glycan 8mer
Structure	
KEGG analogue	KEGG GLYCAN ID: G00245

Sugar residue information

Residue ID	1184655494_03_01	
	Identifier	NAG 1454 B
	IUPAC nomenclature	2-acetamide-2-deoxy-b-D-glucopyranose
	Common name	N-acetyl-b-D-glucosamine
	Anomeric form	b
	Chair conformation	4C1
Residue ID	1184655494_03_02	
	Identifier	NAG 1455 B
	IUPAC nomenclature	2-acetamide-2-deoxy-b-D-glucopyranose
	Comm	N l-b g

Figure 1. An example of glycan structure annotation service. An HTML file containing results of glycan annotation is sent back to the user who submitted a structure file (PDB format). The glycan structure detected in the user-uploaded file is shown in the 'Structure' record and the most similar KEGG GLYCAN entry is shown in the 'KEGG analogue' record. Unmatched substructures (glycoside bonds between mannose residues) are indicated by red question marks.

structure file by getCARBO. As far as we know, there have been no verification services for glycan primary structures. The widely used site, pdb-care (5), a verification service for nomenclature of a saccharide unit, does not have a functionality to verify a glycan primary structure. GDB: Structures provides functionality not only of verifying monosaccharide structures, but also that of glycan primary structures, and will compensate for the lack of tools for modeling biologically meaningful glycan 3D structures.

STRUCTURE AND CONTENT OF GDB:SRUCTURES

GDB:Structures is running on ruby on rails framework with lighttpd web server and MySQL database management system. A glycan annotation program (getCARBO) was written in Java language. A web interface for glycan structure search was developed by using Flash ActionScript 2.0.

All the glycans in GDB:Structures are classified into three types: ligands, *N*-glycans and *O*-glycans. Users can retrieve lists of glycans by specifying the type and length of the glycans. PDB ID is also available to access glycan information. A glycan primary structure search application is also available. Users can draw query glycan primary structures on the Flash web interface and obtain a list of similar glycans in GDB:Structures. The contents of GDB:Structures will be updated regularly, and the statistics of the current content is also presented.

In each glycan information page, the primary structure of the glycan and annotations for the glycan and monosaccharide units are presented. In the case of *N*-glycan, the most similar KEGG GLYCAN entry is also presented. For each monosaccharide unit, the identifier in the PDB, IUPAC nomenclature (9), common name, anomeric form and chair conformation are presented. The KEGG GLYCAN entries are linked to the original data in KEGG (10). The three-letter-identifiers of each monosaccharide unit are linked to Het-PDB navi (11). Each of the PDB entries in GDB:Structures is mutually linked with that in the PDBj (1). Users can interactively browse the 3D structural model of the PDB entry by using a Jmol applet (www.jmol.org). Details of the content of this database are described elsewhere (7).

GLYCAN STRUCTURE ANNOTATION SERVICE

Users can upload their structure file (PDB format) to GDB:Structures to annotate glycan structure(s) in the file by getCARBO. The results of the annotation will be sent back to the users by Email, normally within a few minutes after uploading. The results are summarized in an HTML file (Figure 1). The HTML file presents the results in the same format as the glycan information page of GDB:Structures. In the case of *N*-glycan, the most similar KEGG GLYCAN entry is presented, and unmatched substructures are noted by red question marks. *O*-glycans are not checked in GDB:Structures, because studies on the structures of (especially, non-mammalian) *O*-glycans have not reached the critical mass to perform database searches compared to those on *N*-glycans, of which primary structures and biosynthetic pathways are well known.

CONCLUSION AND FUTURE PERSPECTIVES

GDB:Structures is not only an annotated glycan structure database, but also a web service for verifying the primary structures of glycan 3D structures, which is in high demand by structural biologists (12). This database will help users to determine the characteristics of the glycan structures of their interests and also allow structural biologists to verify their determined glycan structures.

Our *N*-glycan checking process is not always valid and is limited due to the nature of the reference glycan primary structure database. If the glycan primary structure is not found in the reference database, the annotation by getCARBO will point out unmatched substructures on a query glycan structure, even though the 3D structure is built on X-ray diffraction data fine enough to assign unreported motifs. We believe that such incidents should be rare.

The entire data of GDB:Structures will be publicly available following the establishment of an international consensus on the glycan description format (13–15).

ACKNOWLEDGEMENTS

This work was partly supported in part by ‘the National Project on Functional Glycoconjugate Research Aimed at Developing for New Industry’ and ‘Special Coordination Funds for Promoting Science and Technology’ of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government. T.N. thanks Dr Kei Yura at Quantum Bioinformatics Team in Japan Atomic Energy Agency for critical reading of the manuscript. Funding to pay the Open Access publication charges for this article was provided by Special Coordination Funds for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, **10**, 980.
2. Petrescu,A.J., Petrescu,S.M., Dwek,R.A. and Wormald,M.R. (1999) A statistical analysis of *N*- and *O*-glycan linkage conformations from crystallographic data. *Glycobiology*, **9**, 343–352.
3. Wormald,M.R., Petrescu,A.J., Pao,Y.L., Glithero,A., Elliott,T. and Dwek,R.A. (2002) Conformational studies of oligosaccharides and glycopeptides: complementarity of NMR, X-ray crystallography, and molecular modelling. *Chem. Rev.*, **102**, 371–386.
4. Lutteke,T., Frank,M. and von der Lieth,C.W. (2004) Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr. Res.*, **339**, 1015–1020.
5. Lutteke,T. and von der Lieth,C.W. (2004) pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics*, **5**, 69.
6. Crispin,M., Stuart,D.I. and Jones,E.Y. (2007) Building meaningful models of glycoproteins. *Nat. Struct. Mol. Biol.*, **14**, 354.
7. Nakahara,T., Nishimura,S.I. and Shirai,T. (2007) Current aspects of carbohydrate structural bioinformatics. *Curr. Chem. Biol.*, **1**, 265–270.
8. Hashimoto,K., Goto,S., Kawano,S., Aoki-Kinoshita,K.F., Ueda,N., Hamajima,M., Kawasaki,T. and Kanehisa,M. (2006) KEGG as a glycome informatics resource. *Glycobiology*, **16**, 63R–70R.
9. McNaught,A.D. (1996) Nomenclature of carbohydrates (IUPAC recommendations 1996). *Pure Appl. Chem.*, **68**, 1919–2008.
10. Kanehisa,M., Goto,S., Hattori,M., Aoki-Kinoshita,K.F., Itoh,M., Kawashima,S., Katayama,T., Araki,M. and Hirakawa,M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.
11. Yamaguchi,A., Iida,K., Matsui,N., Tomoda,S., Yura,K. and Go,M. (2004) Het-PDB Navi.: a database for protein-small molecule interactions. *J. Biochem.*, **135**, 79–84.
12. Berman,H.M., Henrick,K., Nakamura,H. and Markley,J. (2007) Reply to: building meaningful models of glycoproteins. *Nat. Struct. Mol. Biol.*, **14**, 354–355.

13. Bohne-Lang,A., Lang,E., Forster,T. and von der Lieth,C.W. (2001) LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydr. Res.*, **336**, 1–11.
14. Kikuchi,N., Kameyama,A., Nakaya,S., Ito,H., Sato,T., Shikanai,T., Takahashi,Y. and Narimatsu,H. (2005) The carbohydrate sequence markup language (CabosML): an XML description of carbohydrate structures. *Bioinformatics*, **21**, 1717–1718.
15. Sahoo,S.S., Thomas,C., Sheth,A., Henson,C. and York,W.S. (2005) GLYDE-an expressive XML standard for the representation of glycan structure. *Carbohydr. Res.*, **340**, 2802–2807.